# Blue Gene Project Update

**William R. Pulleyblank**

August 2002

IBM

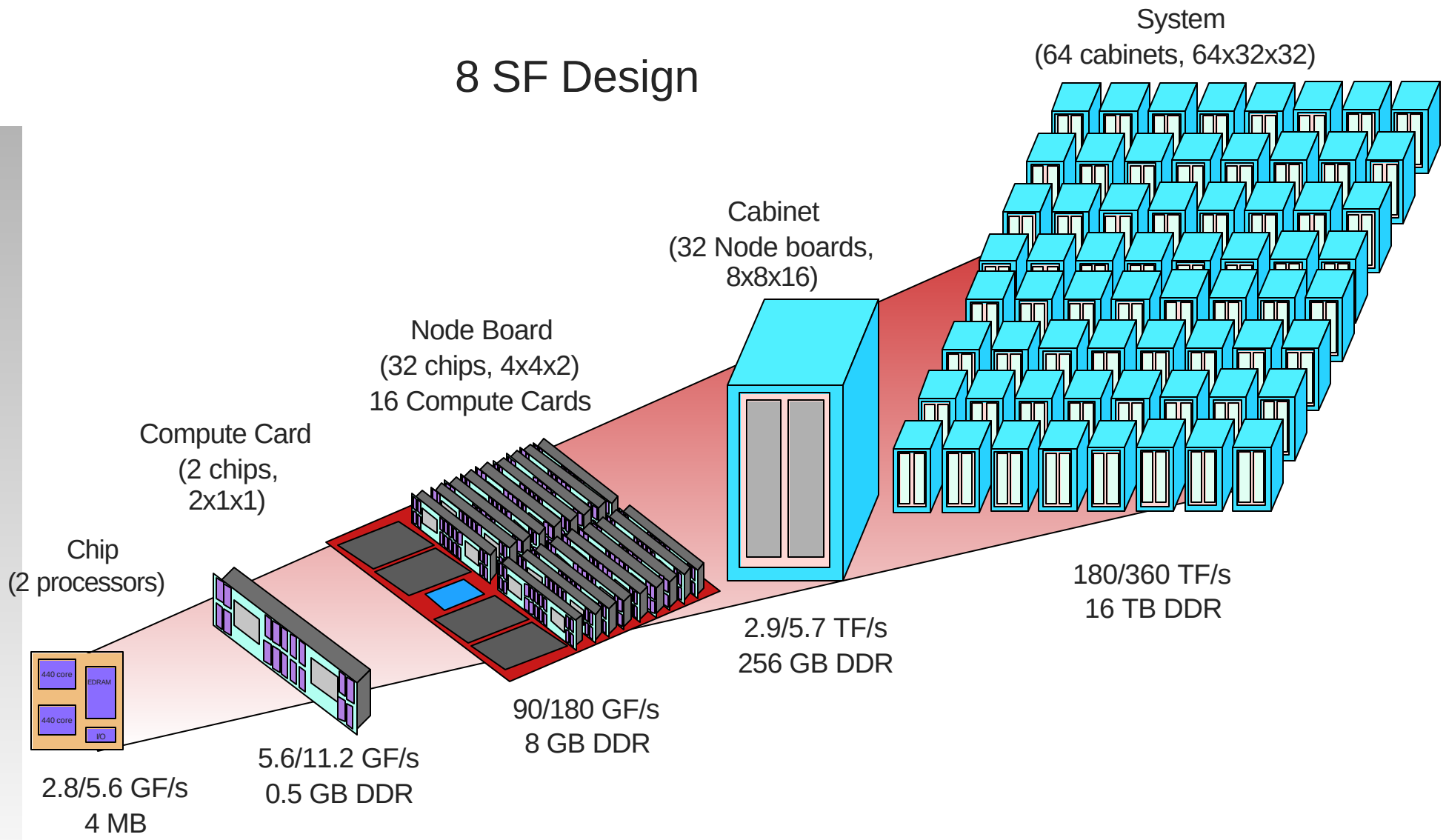
# The Blue Gene Project

- In December 1999, IBM Research announced a 5 year, $100M US, effort to build a petaflop scale supercomputer to attack problems such as protein folding.
- The Blue Gene project has two primary goals:
  - Advance the state of the art of biomolecular simulation.
  - Advance the state of the art in computer design and software for extremely large scale systems.
- In November 2001, a partnership with Lawrence Livermore National Laboratory was announced.

# Blue Gene Project components

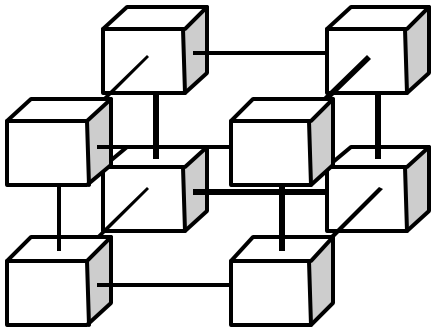
- Two cellular computing architectures
  - Blue Gene/L
  - Blue Gene/C (formerly Cyclops)
  - Blue Gene/D  - variation on BG/L
  - (Blue Gene/P - petaflop machine)
- Software stack
  - Kernels, host, middleware, simulators, OS
  - Self healing, autonomic computing
- Application program
  - Molecular dynamics application software
  - Partnerships, external advisory board

# Blue Gene/L


8 SF Design
System
(64 cabinets, 64x32x32)
Cabinet
(32 Node boards,
8x8x16)
Node Board
(32 chips, 4x4x2)
16 Compute Cards
Compute Card
(2 chips,
2x1x1)
Chip
(2 processors)
440 core
EDRAM
440 core
I/O
2.8/5.6 GF/s
4 MB
5.6/11.2 GF/s
0.5 GB DDR
90/180 GF/s
8 GB DDR
2.9/5.7 TF/s
256 GB DDR
180/360 TF/s
16 TB DDR

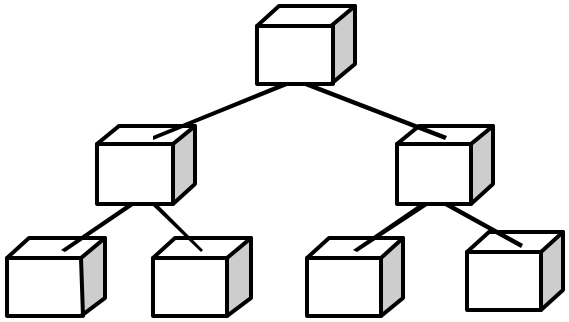
# Blue Gene/L - The Networks

- 65536 nodes interconnected with three integrated networks



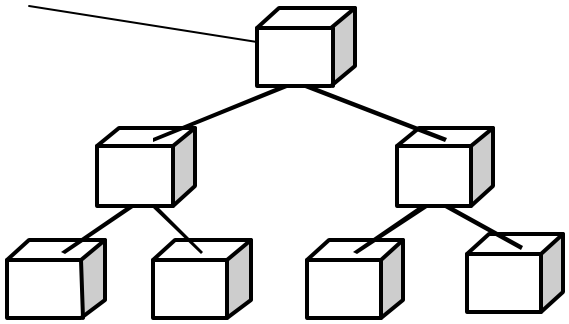
## 3 Dimensional Torus

- Virtual cut-through hardware routing to maximize efficiency
- 1.4 Gb/s on all 12 node links (total of 2.1 GB/s per node)
- Communication backbone
- 67 TB/s total torus interconnect bandwidth



## Global Tree

- One-to-all or all-all broadcast functionality
- Arithmetic operations implemented in tree
- 2.8 Gb/s of bandwidth from any node to all other nodes
- Latency of tree traversal less than 2usec



## Ethernet

- Incorporated into every node ASIC
- Disk I/O
- Host control, booting and diagnostics

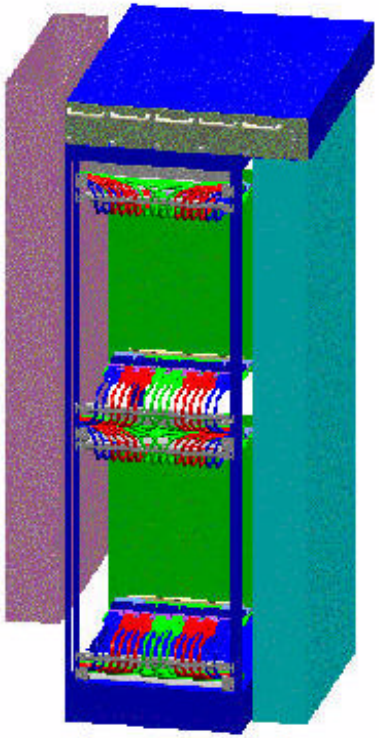
# Blue Gene : a family of systems

- Blue Gene/L
  - Half rack:  512 nodes - 1.5/2.9 TF/s; 128 GB DDR
  - Full rack: 1024 nodes - 2.9/5.7 TF/s; 256 GB DDR
  - ...
  - 64 racks: 64K nodes - 180/360 TF/s; 16 TB DDR
- Blue Gene/D
  - Similar to  BG/L
    - 2.5x DDR
    - higher I/O capability

# Opportunities/Challenges

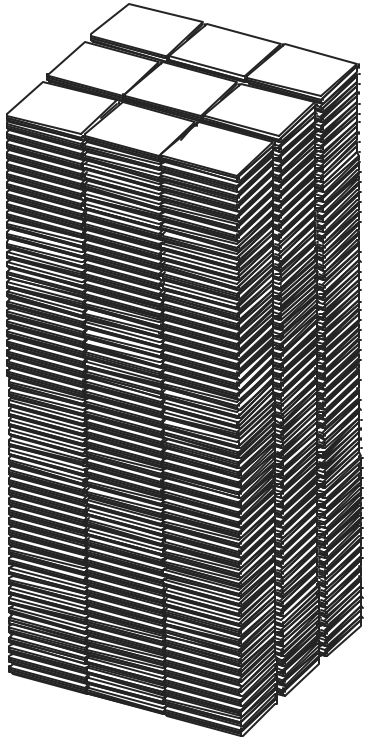
- **Significant opportunities:**
  - Lower price/performance ratios
    - Current SP systems ~ 10,000 $/GFLOPS
    - BG/L ~ 140 $/GFLOPS (hardware cost only)
  - Lower floorspace/performance ratios
    - Cellular architectures use 10-100x less sqft/GFLOPS than current technology
    - 1000 BG/L nodes on a single rack (2000 processors).
  - Lower power/performance ratios
    - < 6 Watts/GFLOPS
- **Challenges:**
  - Distribution of application across partitioned memory
  - Exploit larger compute/ memory ratios than conventional architectures
    - Current technology ~ 1 FLOPS/byte
    - BG/L ~ 10 FLOPS/byte

# System power comparison



BG/L

20.1 kW

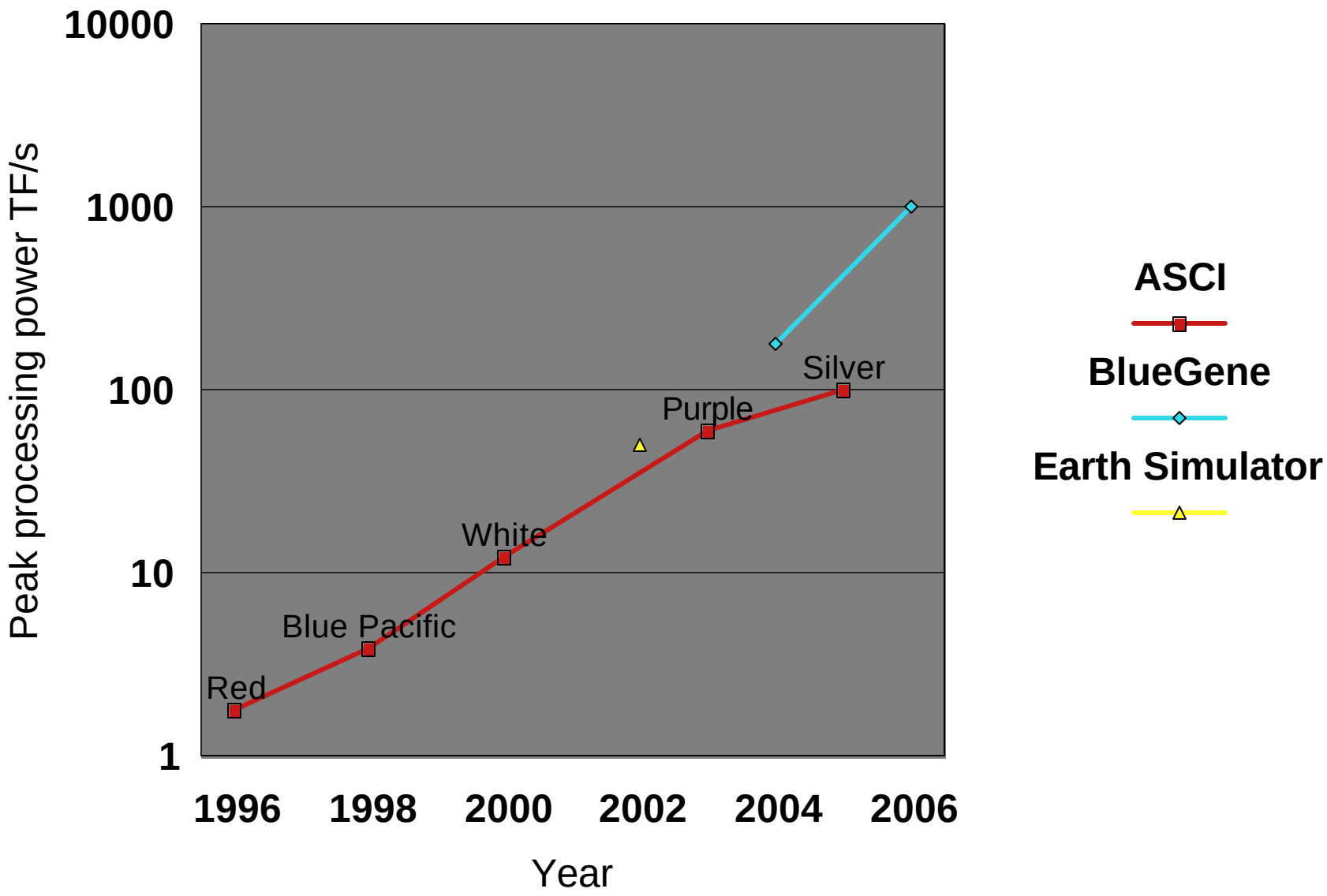


450 Thinkpads

20.3 kW

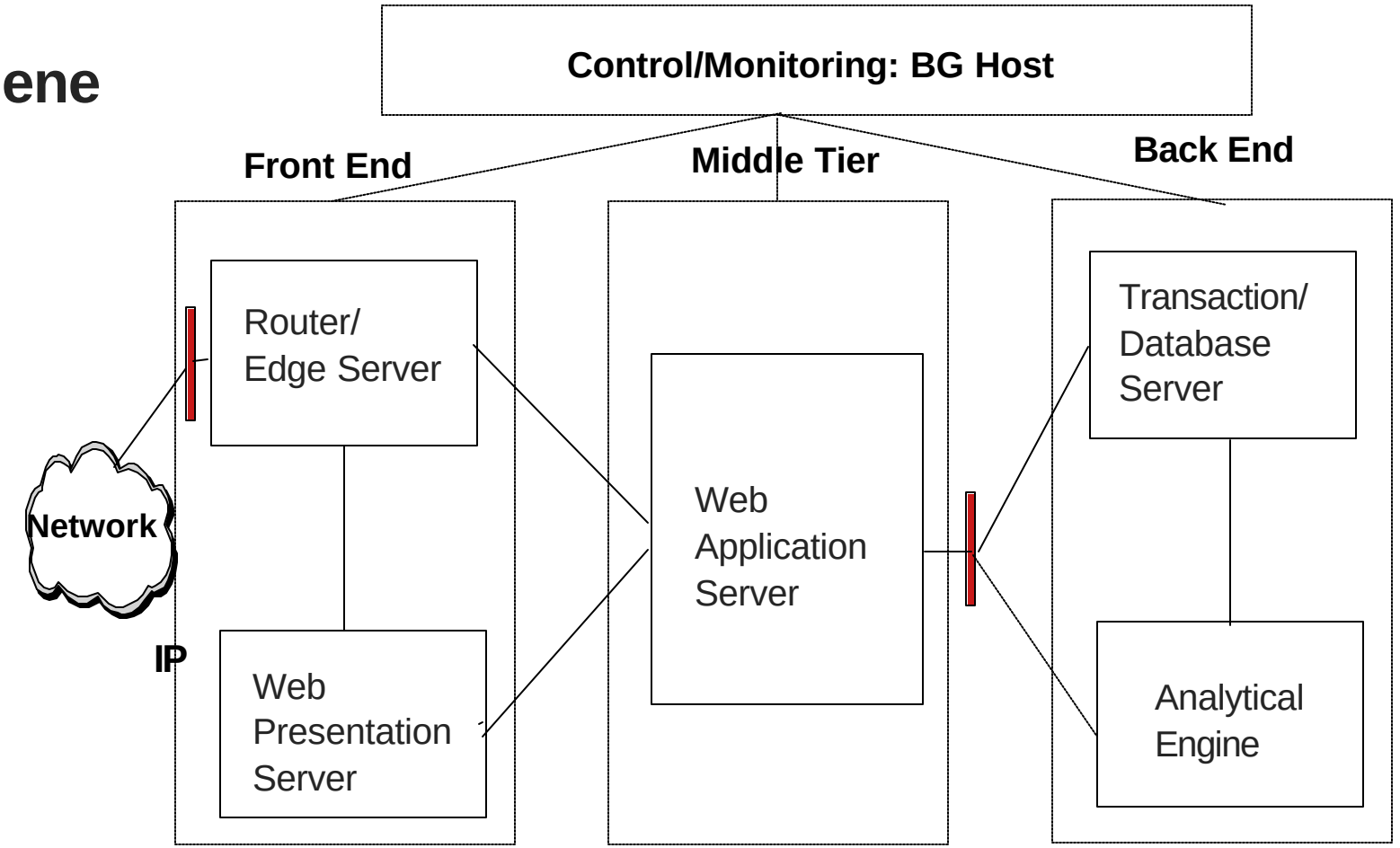
(LS Mok,4/2002)

# HPC Roadmap


10000
1000
100
10
1
Peak processing power TF/s
1996
1998
2000
2002
2004
2006
Year
Red
Blue Pacific
White
Purple
Silver
ASCI
BlueGene
Earth Simulator

# Server Tiers for Business Apps

**Blue Gene**

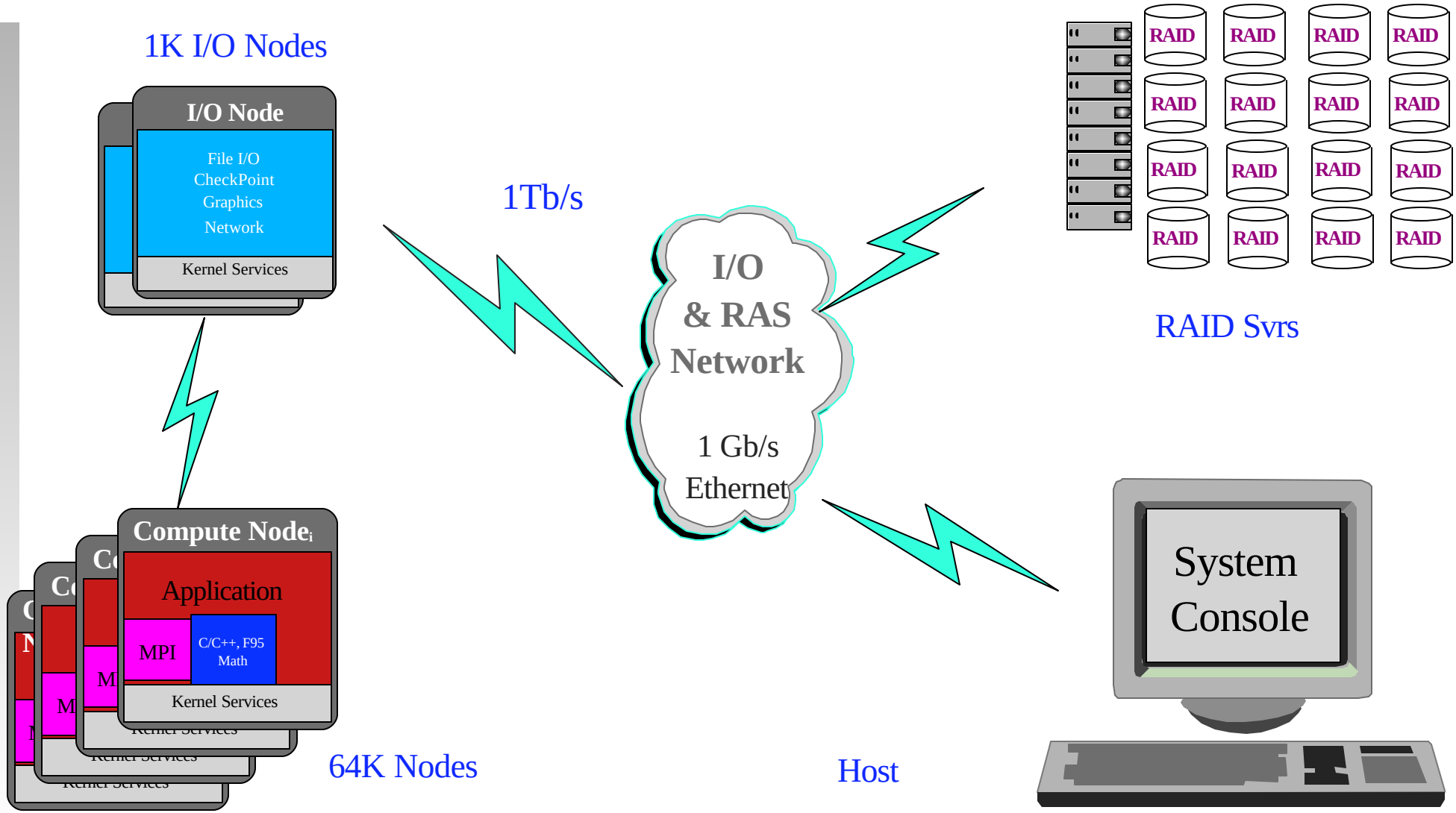



- Run various components (Apache, Websphere, DB2) on different partitions of same BG/L machine.
- Investigate dynamic resource adjustment across partitions.

# System Software Overview

- Operating system - Linux
- Compilers - IBM XL C, C++, Fortran95
- Communication - MPI, TCP/IP, RUDP
- Parallel File System - GPFS, NFS support
- System Management - extensions to CSM
- Job scheduling - based on LoadLeveler
- Math libraries - ESSL
- Simulators
  - Network simulator
  - System level simulator (BLSIM)

# BG/L - Operating Environment



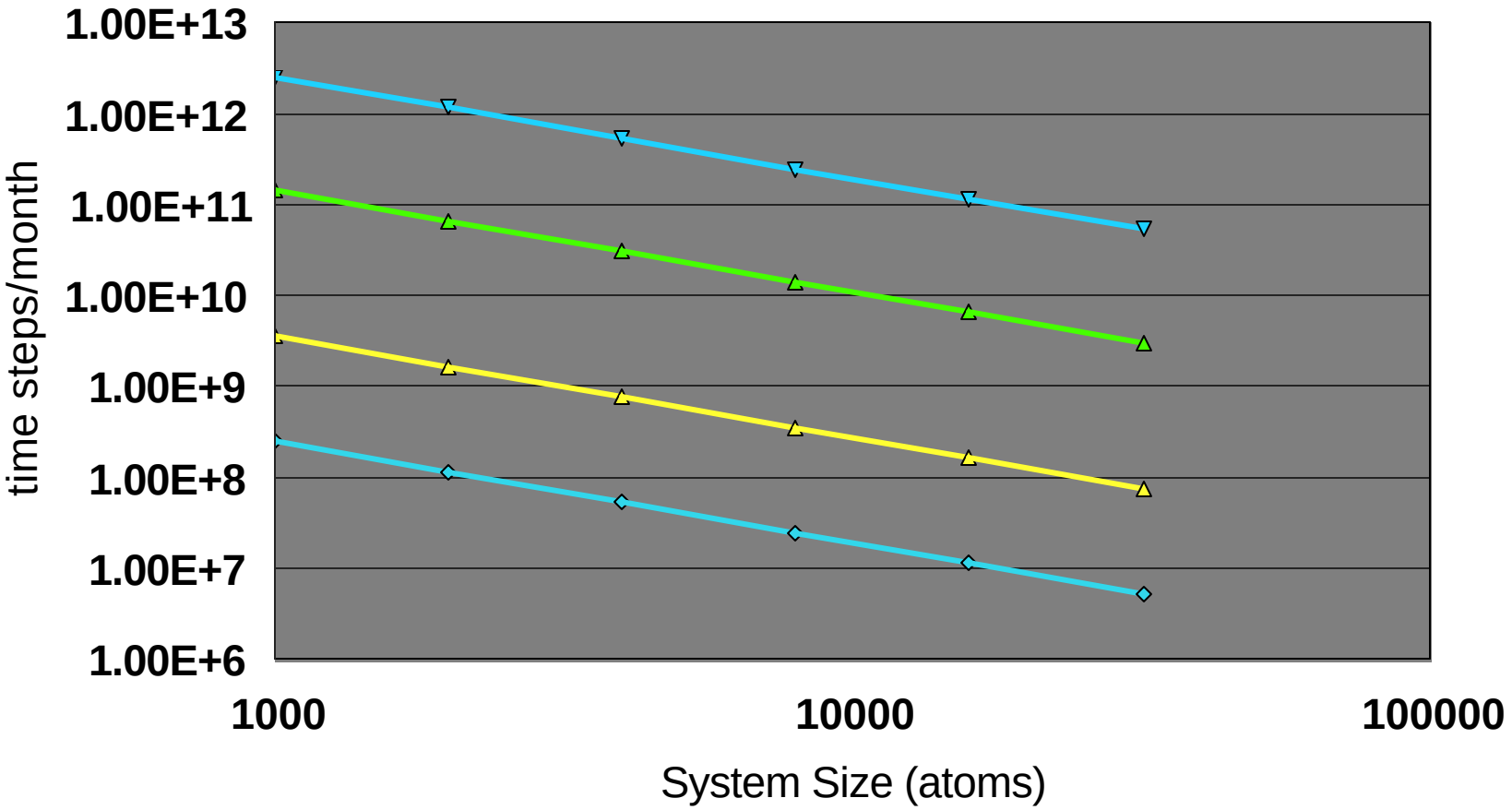
# Blue Matter - a Molecular Dynamics Code

- Separate MD program into three subpackages (offload function to host where possible):
  - MD core engine (massively parallel, minimal in size)
  - Setup programs to setup force field assignments, etc.
  - Analysis Tools to analyze MD trajectories, etc.
- Multiple Force Field Support
  - CHARMM force field (done)
  - OPLS-AA force field (done)
  - AMBER force field (done)
  - Polarizable Force Field (desired)
- Potential Parallelization Strategies
  - Interaction-based
  - Volume-based
  - Atom-based

# Time Scales for Protein Folding Phenomena

| phenomenon | System/size w/solvent | time scale | time step count |
|---|---|---|---|
| beta hairpin kinetics | $\beta$-hairpin/ 4000 atoms | 5$\mu$sec | 10**9 |
| peptide thermo. | $\alpha$-helix, $\beta$-hairpin/400 0 | 0.1-1$\mu$s | 10**8 |
| protein thermo. | 60-100 res./ 20-30,000 | 1-10$\mu$s | 10**9 |
| protein kinetics | 60-100 res./ 20-30,000 | 500$\mu$sec | 10**11 |

# Simulation Capacity



# External Interactions

- System
  - LLNL - all phases
  - Columbia - architecture and system
  - TU Vienna - FFT for BG/L
  - U Barcelona - multithreaded programming models
  - ...
- Science
  - First Blue Gene Protein Science workshop held at San Diego Supercomputer Center, March 2001
  - Second Blue Gene Protein Science workshop held at the Maxwell Institute, U. of Edinburgh, in March 2002
  - Collaborations with ORNL, Columbia, UPenn, Maryland, Stanford, ETH-Zurich, ...
  - Blue Gene seminar series has hosted over 25 speakers at the T.J. Watson Research Center
  - Blue Gene Applications Advisory Board formed with 15 members from the external scientific and HPC communities.
  - ...

# Blue Gene Project Update

**William R. Pulleyblank**

August 2002

IBM